Le système de correspondance phonétique dans le dictionnaire étymologique: „Les éléments finno-ougriens dans le vocabulaire du hongrois"*

1. À fin de représenter plus clairement l'identité étymologique des mots hongrois avec des mots correspondants des langues apparentées, le dictionnaire étymologique portant le titre „Les éléments finno-ougriens dans le vocabulaire du hongrois" indique, entre autres, la forme phonétique reconstituée de l'antécédent de la famille de mots supposé appartenir à l'ouralien, au finno-ougrien ou ougrien primitifs. Sous forme de symboles concis, ces formes primitives servent à formuler le résultat que l'on obtient en tirant les conséquences possibles des correspondances entre les langues apparentées.

Dans les deux volumes du dictionnaire déjà parus, on peut trouver 336 formes primitives ayant à l'intérieur un élément consonantique (consonne simple, -géminée, groupe de consonnes composé de deux ou, plus rarement, de trois consonnes). En comparant les représentations phonétiques du hongrois actuel avec leurs antécédents supposés appartenir à la langue primitive, on aboutit à un système compliqué, divergeant et en même temps entrelacé des correspondances phonétiques.

Ce système des correspondances phonétiques peut être analysé, du point de vue de la théorie de l'information, comme système statistiquement indépendant du temps, dont les caractéristiques sont déterminées par la fréquence des éléments qui y figurent. Autant que je sache, les relations de fréquence internes, c'est á dire les lois statistiques indiquant les correspondances phonétiques (ces dernières appliquées au cours des recherches étymologiques ouraliennes) n'ont pas encore été traitées. Pourtant, de semblables recherches pourraient contribuer à l'exploration des indices quantitatifs caractérisant tout le système, ce qui permettrait de mesurer le degré d'authenticité des types établis des correspondances phonétiques. C'est ce que j'essaie de démontrer en faisant des recherches sur un sous-système du système de correspondances phonétiques traité dans le dictionnaire, qui inclue deux tiers des correspondances consonantiques á l'intérieur du mot.

2. En partant de l'état actuel des langues apparentées pour explorer l'antécédent le plus possible, la mise en correspondance des sons a une structure et une démarche logique de sens invers par rapport aux processus réels de l'évolution de la langue. Donc le variable d'état « input « n'est autre chose que la représentation phonétique de la langue dérivée (en ce cas le hongrois), tandis que les signes «output« sont les sons hypothétiques de la langue primi-

* A magyar szókészlet finnugor elemei. I. Budapest 1967., II. Budapest 1971.

tive. La structure même du système est déterminée par le réseau de la rélation des signes input et output, ainsi que par les fréquences relatives à cette relation.

Le degré d'authenticité des correspondances phonétiques est détermin par les facteurs quantitatifs suivants:

1. Par la fréquence des représentations phonétiques input (c'est à dire par le pourcentage des représentations consonantiques — consonnes simples et géminées ou groupes de consonnes — dans l'ensemble des 336 mots analysés).
2. Par le nombre des représentations phonétiques output (cette fois par le nombre des représentations primitives possibles) correspondant à chaque signe input.
3. Par la répartition de la charge des types divers de correspondance phonétique, c'est à dire par la dispersion de chaque représentation input entre ses représentations output.

De ce qui suit, ces trois critères seront nommées «fréquence«, «nombre d'antécédents« et «indice de dispersion«.

3. Le premier supplément offre l'illustration de la répartition réelle des correspondances phonétiques. Les 336 pairs correspondants s'arrangent dans 118 types de correspondance d'après leurs signes input et output qui doivent différer au moins dans l'une des marques significatives. Le mombre des représentations phonétiques input est 39. On en peut distinguer 4 sous-groupes:

1. Consonnes simples: 20. Absence de *C*, *DZS*, *ZS*, *F* et de *TY*.
2. Consonnes géminées: les six suivantes; *NNY*, *TTY*, *GGY*, *GG*, *SSZ* et *LL*.
3. Groupes de consonnes: 12 groupes bipartits. Les groupes caractéristiques en sont les suivants (selon la présence ou l'absence d'un élément occlusif):
   type 01: nasale, fricative ou liquide + occlusive
   MB, *NG*, *NGY*, *JD* et *LGY*.
   type 00: fricative, liquide ou trémulante + nasale ou fricative
   JSZ, *LH*, *RM*, *RNY*, *RV*, *RS*.
   type 10: occlusive + nasale
   GYM
   type 11: occlusive + occlusive
   ce type n'est pas représenté.
4. Ø (manque de son): Toutes les représentations phonétiques de degré Ø dans la forme lexicographique du mot, même s' il y a des altérnances thématiques où quelque consonne (*V*, *J*, *H*) est présente, appartiennent à ce groupe.

Il y a en somme 75 sortes de représentations phonétiques output. À cause de l' incertitude de la mise en correspondance des sons, on a obtenu 21 signes output doubles (donc deux antécédents pour un signe input) et un signe triple. Entre les 20 consonnes supposées de la langue primitive il y a 18 qui sont des représentations consonantiques simples; le **š* apparait dans une seule position, dans le groupe de consonnes **kš*, alors que**γ* apparait ou bien dans des groupes de consonnes, ou bien comme variante facultative près de **k* et de **w*. Il y a trois consonnes géminées: **pp*, **tt*, **kk*; 40 groupes de consonnes doubles et

trois groupes de consonnes triples. La plupart des groupes de consonnes doubles est de type 00 (23 groupes) il y a 11 groupes de type 01 et il n'y a que trois groupes qui sont de type 10 (**kš*, **ks* et **ps*). Il y a également trois groupes de type 11: **kt*, **tk* et **pt*. Il semble que les groupes formés de trois consonnes peuvent être décomposés en deux sous-groupes: en 2 + 1 (**ńć|k*) et en 1+2 (**η|ks* et **w|kk*).

4. Les valeurs de fréquence des signes input se trouvent dans la première colonne du deuxième supplément. La distribution uniforme des 336 données entre les 39 représentations phonétiques donnerait une fréquence moyenne de 8,6. La fréquence de *R* est le quintuple, celle de *L* et de Ø est le quadruple et celle de *GY*, de *Z*, de *J*, de *K*, de *G* et de *P* est le double de cette valeur moyenne. En revanche il y a 13 entre eux (*TTY*, *DZ*, *H* et presque tous les groupes de consonnes) qui n'apparaissent qu'une fois.

Les nombres d'antécédents primitifs se trouvent dans la deuxième colonne du 2ème supplément. Les 39 représentations phonétiques input ont en somme 118 antécédents possibles. Le nombre d'antécédents correspondant à chaque représentation phonétique est en moyenne 3,0. *GY*, Ø et *L* ont un très grand nombre d'antécédents: plus que le triple; *Z*, *J*, *R*, *S* et *LL* ont un nombre d'antécédents près du double de la valeur moyenne; *T*, *SZ*, *M*, *NY*, *D* et *CS* ont un nombre d'antécédents proche de la valeur moyenne; *P*, *G*, *N* et *B* se distinguent des autres par leur nombre d'antécédents très petit par rapport à leur assez grande fréquence.

Sur la base de la proportion de charge des divers types de correspondance phonétique appartenant à chaque représentation phonétique input et déterminé par le nombre d'antécédents, nous pouvons obtenir les indices de dispersion par la méthode suivante:

Si la représentation phonétique n'a qu'un seul antécédent, la mise en correspondance a une univocité de 100%, donc l'indice de dispersion est 1,000. Mais si le signe input peut avoir deux ou plus de deux outputs possibles, la mise en correspondance aura certainement une dispersion positive: elle sera inférieure à un et sa valeur actuelle dépandra toujours de la fréquence du signe input et des proportions de distribution appartenant à chaque signe output.

La formule générale pour cette fonction est la suivante: $D = \sum_{i=1}^{n} \left(\frac{k_i}{B}\right)^2$

où D = indice de dispersion
B = fréquence input
k_i = valeur de charge pour chaque signe output $k_1, k_2 \ldots k_i \ldots k_n$
n = nombre des signes output

En traduisant cette formule au language des calculs pratiques; nous devons faire l'addition des carrés des fréquences pour chaque output et diviser la somme obtenue par le carré de la fréquence input.

Voici un exemple concret de ce procédé. Supposons que la fréquence d'un signe input est 4. Dans ce cas, le nombre des signes output ne peut être en principe que 1, 2, 3 ou 4.

Si le nombre des outputs est *1* il n'y a pas de dispersion, $D = 1{,}000$.

Si le nombre des outputs est *2* les proportions de distribution possibles sont 1:3 et 2:2

si nous avons 1:3, $D = \frac{1^2 + 3^2}{4^2} = \frac{10}{16} = 0{,}625.$

si nous avons 2:2, $D = \frac{2^2 + 2^2}{4^2} = \frac{8}{16} = 0{,}500.$

Si le nombre des outputs est *3* la proportion de distribution ne peut être que 1:1:2

$$D = \frac{1^2 + 1^2 + 2^2}{4^2} = \frac{6}{16} = 0{,}375.$$

Si le nombre des outputs est *4* la proportion de distribution ne peut être que 1:1:1:1

$$D = \frac{1^2 + 1^2 + 1^2 + 1^2}{4^2} = \frac{4}{16} = 0{,}250.$$

Les résultats des calculs pour les indices de dispersion se trouvent dans la 3ème colonne du 2ème supplément. Mais la dispersion seule ne peut pas servir de mesure d'authenticité pour les correspondances phonétiques. Car il est évident que, même avec des proportions de distribution output identiques, les fréquences des représentations phonétiques input peuvent être très différentes: par exemple les fréquences input 4 et 40 ayant une proportion de distribution 1:3 ont également un indice de dispersion 0,625 ce qui signifie dans le cas de la première 1 et 3 apparitions, tandis que dans le cas de la deuxième 10 et 30. Donc les facteurs de fréquence jouent aussi un rôle important dans le jugement du degré d'authenticité.

Les valeurs qui servent à démontrer le degré d'authenticité des correspondances phonétiques (faute de meilleur terme on les nommera «indices d'authenticité«) peuvent être obtenu par la multiplication des indices de dispersion par les fréquences input (v. les résultats dans la 4ème colonne du 2ème supplément).

Avec des dispersions identiques, les écarts (réduits ou augmentés) de fréquence des correspondances phonétiques motivent l'écart (réduit ou augmenté) de l'indice d'authenticité, qui est proportionnel et conforme aux indices de dispersion.

Avec des fréquences identiques, les écarts (réduits ou augmentés) dans la proportion de distribution motivent l'écart de sens invers des indices de dispersion et le changement des indices d'authenticité proportionnel á l'écart des indices de dispersion.

Nous pouvons donc constater que la fréquence est directement proportionnelle á l'authenticité, tandis que la dispersion y est inversement proportionnelle. Ceci est conforme à la thèse des étymologues, selon laquelle plus une correspondance phonétique a des pairs correspondant univoques, parmi des représentations phonétiques limitées autant que possible, plus elle est authentique.

5. Quelles conséquences peut-on tirer des calculs statistiques en ce qui concerne l'authenticité des correspondances phonétiques? Les lois statistiques qui déterminent le sous-système des correspondances phonétiques peuvent

être rendues visibles par la représentation graphique des paramètres internes, statistiques de ces correspondances (v. 3ème supplément). On y voit clairement que 5 catégories des correspondances phonétiques peuvent être distinguées sur la base des indices d'authenticité:

1. La première catégorie ne contient qu'une seule représentation phonétique, notamment *R*. **r* primitif se garde avec grande fréquence dans le hongrois, mais **k*, **w*, **j* et **γ* primitifs peuvent disparaître de son environnement.
2. *L*, *P* et *K* peuvent figurer dans cette catégorie. **l* primitif garde sa grande fréquence dans le hongrois, mais **k*, **w*, **j* et **γ* primitifs peuvent disparaître de son environnement, ces derniers montrant une plus grande fréquence dans cette position que dans l'environnement de **r*. Antécédents primitifs de *L* peuvent également être **δ* et plus rarement **wδ* ou **nt*. Malgré leurs fréquences presque identiques, l'indice d'authenticité de *L* est donc moins que la moitié de celui de *R*. Dans ce sous-système, c'est *P* qui a le correspondant le plus univoque: tous ses antécédents sont déductibles de **pp*. *K* aurait une pareille position, s'il n'avait pas, comme conséquence de l'influence modifiante des correspondances des langues apparentées, deux fois **wkk* au lieu de **kk* comme antécédent primitif. Donc *K* montre, malgré sa plus grande fréquence, un degré d'authenticité plus bas. Quant à *T*, on pourrait demander pourquoi il ne figure pas entre les autres occlusives *P* et *K*. La fréquence de *T* n'est pas moins élevée que celle de *P* et de *K*, mais à cause de son indice d'authenticité plus bas, il a trouvé place de deux catégories plus loin. C'est une conséquence de sa proportion de distribution très grande: il a comme antécédents **tt*, **kt*, **tk* et **pt*, donc tous les groupes: occlusive + +occlusive. Cette incertitude dans les correspondances de *T* est, jusqu'à nos jours, la source de nombreux points d'interrogation.
3. Dans cette catégorie nous avons déjà 6 représentations phonétiques diverses. *G* a deux antécédents primitifs: **ηk* et **η*. Ce qui est étonnant, c'est que le dernier est beaucoup plus chargé (11:4). La correspondance univoque de N est variée par l'antécédent **m* dans *indul*. Mais cette alternance de la consonne motivée par le suffixe ne fait écarter l'indice d'authenticité que de 0,8. L'influence modifiante des correspondances sporadiques dans le cas de *Z* est d'autant plus grande. Sur la base de ses antécédents **t* avec un indice d'authenticité de 11,0, il devrait figurer dans la 2ème catégorie, mais la dispersion des correspondants mène à une réduction de 3,8. Dans le cas de *D* c'est pareil: ici, c'était la dispersion des correspondants **nt* et **mt* qui a causé la réduction de valeur. La valeur d'authenticité de *M* est réduite par le fait que **l*, **l'* ou **δ'* peuvent disparaître de son environnement, ce qui rend ses correspondances plus incertaines. Ø (manque de son) a une fréquence et une dispersion également grandes. C'est grâce à ces deux facteurs que son indice d'authenticité approche de la moyenne de tout le système: même arrondi à deux chiffres, cela donne 5,8. Ø a 11 antécédents divers qui ont la caractéristique commune suivante: dans la plupart des cas, ils se trouvent près de *V*, *J* ou plus rarement de *H*: sons remplissant des hiatus. Conformé-

1^er Supplément: Répartition des correspondances phonétiques

M	*m* 7	*lm* 1	*ƚm—δ'm* 1								
N	*n* 9	*m* 1									
NY	*n* 5	*ń* 4	*δ'* 1								
NNY	*n* 1	*ń* 2									
P	*pp* 14										
T	*tt* 4	*kt* 4	*tk* 1	*pt* 1							
TTY	*tt* 1										
K	*kk* 13	*wkk* 2									
B	*mp* 4										
D	*nt* 8	*mt* 1	*mt—nt* 1								
GY	*ńć* 5	*ć* 2	*ńć—ć* 2	*č* 2	*j* 2	*jη* 1	*ƚ* 1	*ƚ—δ'* 2	*δ'* 1	*l* 1	*n* 1
GGY	*ƚ* 1	*ƚ—δ'* 1									
G	*η* 11	*ηk* 4									
GG	*η* 1	*ηk* 1									
CS	*č* 1	*ć* 1	*ćk* 1								
DZ	*tt—t* 1										
SZ	*ś* 5	*ć* 2	*ć—ś* 1	*s* 1							
SSZ	*ć* 1	*ńć—ć* 1									
S	*ć* 6	*č* 2	*ńć—ć* 2	*č—ć* 1	*ćk* 1	*śk* 1					
H	*kš* 1										
V	*p* 2	*η* 1									
Z	*t* 11	*č* 2	*ć* 1	*ć—ś* 1	*ś* 1	*s* 1	*ps* 1				
J, LY	*j* 8	*δ'* 2	*l* 2	*η* 1	*jw* 1	*ľw* 1	*j—jγ* 1				

L	*l* 24	*δ'* 5	*lk* 4	*lγ* 1	*lj* 1	*wl* 1	*l—lk* 1	*l—lj—lk* 1	*wδ* 1	*nt* 1	
LL	*1* 4	*lk* 3	*1γ* 1	*lk—lγ* 1	*ηl—lη* 1						
R	*r* 36	*rk* 3	*r—rk* 1	*r—rw* 1	*r—rj* 1	*rw—rγ* 1					
Ø	*j* 11	*m* 5	*η* 5	*k* 5	*w* 3	*w—γ* 3	*w—η* 1	*k—γ* 1	*ks* 2	*jk* 1	*ηks* 1
MB	*mp* 1										
NGY	*ń* 1	*ńć—ć* 1									
NG	*n* 1										
GYM	*ćm* 1										
JD	*δ't* 1										
JSZ	*jć* 1										
LGY	*δ'w* 1										
LH	*lw—lη* 1										
RM	*lm* 1										
RNY	*rń* 1										
RV	*rp* 1	*rp—rw* 1									
RS	*ńćk—ćk* 1										

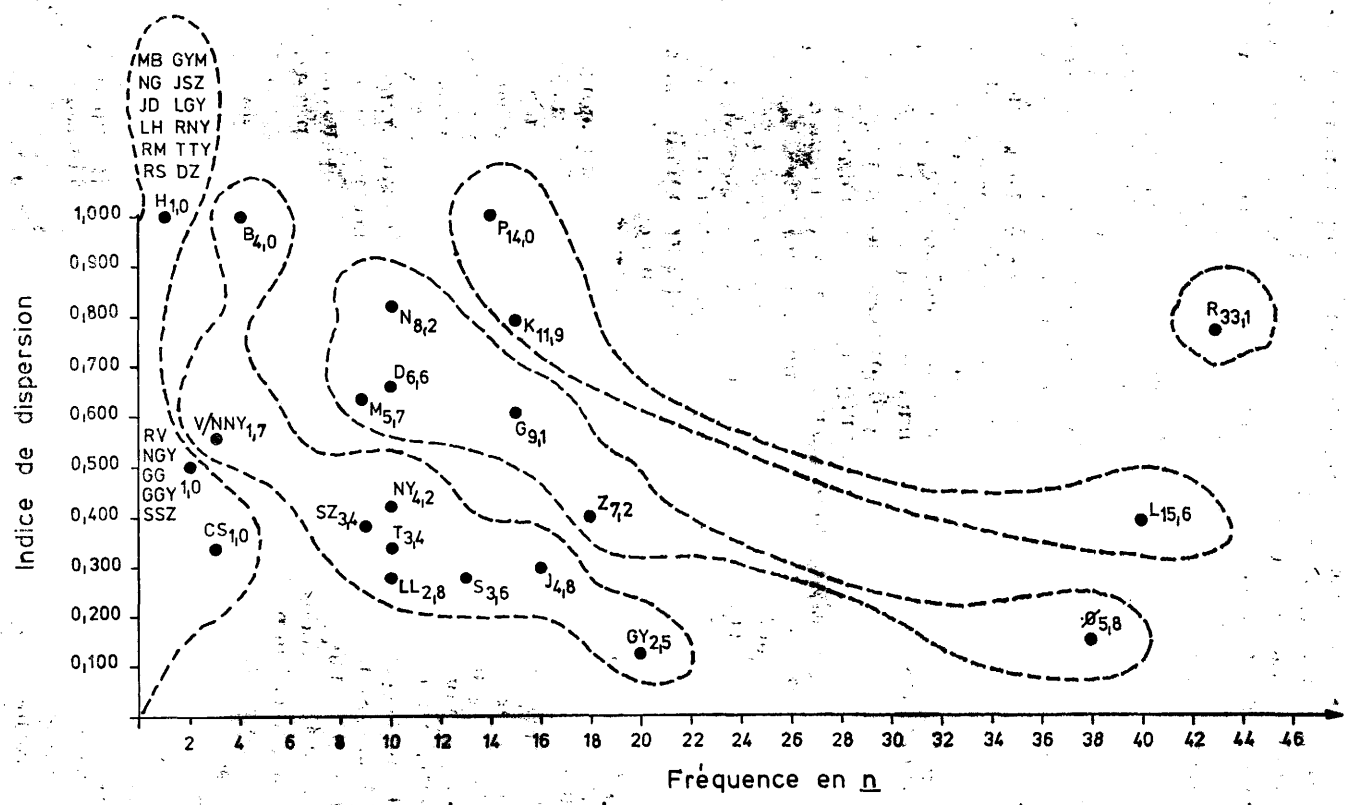
2ème Supplément: Planches statistiques

Signe input	Fréquence	Nombre d'antécédents	Indice de dispersion	Indice d'authenticité
M	9	3	0,633	5,7
N	10	2	0,820	8,2
NY	10	3	0,420	4,2
NNY	3	2	0,556	1,7
P	14	1	1,000	14,0
T	10	4	0,340	3,4
TTY	1	1	1,000	1,0
K	15	2	0,791	11,9
B	4	1	1,000	4,0
D	10	3	0,660	6,6
GY	20	11	0,125	2,5
GGY	2	2	0,500	1,0
G	15	2	0,609	9,1
GG	2	2	0,500	1,0
CS	3	3	0,333	1,0
DZ	1	1	1,000	1,0
SZ	9	4	0,382	3,4
SSZ	2	2	0,500	1,0
S	13	6	0,279	3,6
H	1	1	1,000	1,0
V	3	2	0,556	1,7
Z	18	7	0,401	7,2
J, LY	16	7	0,297	4,8
L	40	10	0,390	15,6
LL	10	5	0,280	2,8
R	43	6	0,771	33,1
ø	38	11	0,154	5,8
MB	1	1	1,000	1,0
NGY	2	2	0,500	1,0
NG	1	1	1,000	1,0
GYM	1	1	1,000	1,0
JD	1	1	1,000	1,0
JSZ	1	1	1i000	1,0
LGY	1	1	1,000	1,0
LH	1	1	1,000	1,0
RM	1	1	1,000	1,0
RNY	1	1	1,000	1,0
RV	2	2	0,500	1,0
RS	1	1	1,000	1,0
Moyenne	8,6	3,0	0,674	5,8

ment á ceci, on pourrait énumérer ici les correspondances *V* — **p* ou *V* — **η*, *J* — **η* et *H* — **kš*. Mais cette hétérogénéité des correspondances nous conseille d'être prudent quant au jugement des antécédents possibles de la manque de son dans les mots hongrois.

4. Parmi les 10 représentations phonétiques de la 4ème catégorie, c'est *B* dont les correspondants sont tout à fait univoques; son bas degré d'authenticité est la conséquence de son apparition rare. Malgré sa plus grande fréquence, *NY* doit également figurer dans cette catégorie à cause de la dispersion de ses antécédents entre **ń* et **n*. Comme c'est **n* qui domine entre les deux, il faut donc compter avec une palatalisation secondaire. La correspondance sporadique avec **δ'* dans *enyv* modifie l'indice d'authenticité de *NY* seulement de 0,4. Quant à *J*, on n'a pas distingué les graphèmes *j* et *ly*: mais leur distinction n'aurait pas changé le position de *J*; pour deux antécédents de moins, il y aurait une fréquence de trois plus basse, donc l'hétérogénéité des correspondants de *J* ne changerait pas. La dispersion des correspondants de *SZ* est une conséquence de la présence de **ć* alternant avec **ś* et une correspondance irrégulière ,sporadique de **s*. Il est intéressant de voir que c'est **ć* qui domine parmi les antécédents de *S*; sa valeur est donc plus basse à cause des autres correspondances. *GY* semble être une sorte de réceptacle des antécédents primitifs divers: c'est lui qui a la plus grande dispersion dans tout le système. Par conséquence, son indice d'authenticité est très petit malgré la fréquence élevée qu'il a. Les deux domaines principaux de ses correspondances sont d'une part les affriquées (**ć*, **č* et surtout **ńć*) d'autre part les fricatives et les liquides palatales ou palatalisées (**j*, **δ'*, **l'*). En dehors de cela, et grâce à certaines palatalisations secondaires sporadiques, **l* et **n* appartiennent aussi au «champ d'attraction« de *GY*. L'apparition des consonnes longues dans la 4ème et la 5ème catégories mériterait une analyse à part mais je me contenterai de remarquer que l'allongement peut être expliqué seulement en partie par la supposition d'un groupe de consonnes comme antécédent; dans la plupart des cas, la raison de l'allongement est à chercher ailleurs. Il ne semble pas forcé de chercher des relations entre les mots *faggyú*, *hattyú*, *könnyű*, *hosszú* et peut être *holló* et *messze*. Quant aux mots *meggy*, *menny* et *könny* (arch. *könyü*, *könyv*) c'était peut être la volonté d'éviter l'identité phonétique avec des mots *megy*, *meny*, *könyv* qui a pu jouer un rôle.
5. Des cas individuels et des apparitions uniques sont caractéristiques à la 5ème catégorie. Les groupes de consonnes plus rares en constituent une partie (*GYM*, *JSZ*, *RNY*, *RV*), tandis que les correspondances les plus problématiques s'y trouvent aussi: des mots comme *méh*, *felhő*, *edz*, *domb*. On y trouve également des groupes de sons comportant des consonnes «non-étymologiques«.

6. À cause des possibilités techniques limitées, ces recherches ne pouvaient prendre en considération qu'une partie du système de correspondances phonétiques traitées dans le dictionnaire. Une analyse pareille de tout le système pourrait certainement donner davantage des conclusions utiles.



3ème Supplément: Représentation graphique des paramètres statistiques